In *Maximum Entropy and Bayesian Methods*,
K.M. Hanson and R.N. Silver, eds. (Kluwer, 1995, to be published)

TOWARD OPTIMAL OBSERVER PERFORMANCE OF DETECTION AND DISCRIMINATION TASKS ON RECONSTRUCTIONS FROM SPARSE DATA

R.F. WAGNER, K.J. MYERS, D.G. BROWN AND M.P. ANDERSON
Center for Devices & Radiological Health, Rockville MD 20857

AND

K.M. HANSON
Los Alamos National Laboratory, Los Alamos NM 85745

Abstract. It is well known that image assessment is task dependent. This is demonstrated in the context of images reconstructed from sparse data using MEMSYS3. We demonstrate that the problem of determining the regularization- or hyperparameter α has a task-dependent character independent of whether the images are viewed by human observers or by classical or neural-net classifiers. This issue is not addressed by Bayesian image analysts. We suggest, however, that knowledge of the task, or the use to which the images are to be put, is a form of prior knowledge that should be incorporated into a Bayesian analysis. We sketch a frequentist approach that may serve as a guide to a Bayesian solution.

Key words: observer performance, tomographic reconstruction, MEMSYS, hyperparameter selection

1. Introduction

Images are generally produced for the purpose of performing visual tasks. A visual task typically consists of an interpretation of an image and a decision about its content. The major premise in the field of image assessment is that the ranking of imaging systems is *task dependent*. This point was demonstrated in the literature about twenty-five years ago [1] and a consensus on the issue developed almost immediately. Nevertheless, the point is rarely addressed in the Bayesian and Maximum Entropy communities. It is the purpose of the present work to demonstrate the task dependence of image assessment in the context of image reconstruction from sparse data using a maximum entropy method. We shall see that the search for an optimal regularization parameter in this method does not have a unique solution: the solution depends on the task. We shall compare frequentist and Bayesian

approaches to the issue and will offer a frequentist approach to the problem that may serve as a guide to a Bayesian solution.

2. Tasks and observers

We consider two very broad categories of imaging tasks: lesion or target detection, and target discrimination or classification. A simple example of a detection task in medical imaging is the task of determining whether a site in an image contains a lesion or is only representative of normal background. A simple example of a discrimination task is the task of viewing a blood vessel and determining whether it has significant narrowing (stenosis) or not. For the special case where an imaging system is linear and shift-invariant, the detection task can be considered as a task that is concentrated in the low spatial frequencies. The discrimination task, however, requires no low spatial-frequency information: it is a mid- to high-frequency task.

In the field of image assessment a number of classes of *image observers* are considered. An observer is defined in terms of how the task is implemented or performed. Here we limit ourselves to binary tasks and the case where the signal is known exactly (SKE). In this paradigm the observer is focused on the region of interest and has to decide which of two hypothesized states gave rise to the data, e.g., lesion present, or only background present. (Performance of some complex tasks is related to performance of SKE tasks in Refs. 2 and 3.) The most common observer is the human, but there are other ways of realizing an observer using a machine or computer algorithm. These include various matched filters, classical decision rules based on the likelihood function (the foundation for the matched filters [4]), the "proper" Bayesian rule based on the ratio of posterior proabilities for the two hypotheses given the data [5], and a growing number of classical and neural-net extensions of these decision rules.

Once a task and an observer are defined, standard methods from statistical decision analysis can be applied to assess the performance of the task by the observer. In the present work we investigated the performance of human and machine observers on lesion detection and classification tasks as a function of the regularization parameter in the reconstruction algorithm.

3. Image "acquisition" and reconstruction

The present work is an extension and interpretation of previous work in this series that considered limited angle tomography (eight views over 180 degrees). The data are derived from a simulation of parallel-projection image acquisitions, 128 samples per projection, and additive Gaussian noise. Images reconstructed from such data will be corrupted by artifacts due to the small number of projections or views, as well as by the additive measurement or detection noise. For additional details, see [5-7].

We concentrate here on MEMSYS3, a Bayesian method of regularized reconstruction due to the Cambridge school of Gull and Skilling [8,9]. The algorithm

includes a numerically efficient method for minimizing a functional F, where

$$F = \tfrac{1}{2}\chi^2 - \alpha S \ . \tag{1}$$

The term involving χ^2 is the usual measure of misfit between the data and the reconstruction $\hat{f}$ (referred to the data domain). The quantity S in the second term is the Cambridge adaptation of the Shannon entropy ($-\sum_i \hat{f}_i \ln \hat{f}_i$). The second term may thus be thought of as the exponent in an entropic prior probability distribution or, rather, a family of entropic prior probability distributions in the (hyper-) parameter α. The parameter α may be thought of as a regularization parameter that determines the degree of smoothness in the final reconstruction. We shall now give several views on the determination of the value of α.

4. Frequentist and Bayesian perspectives

In classical statistics, or the so-called frequentist perspective, one is interested in the long-term average behavior of a method of estimation or inference. In the context of image assessment, one assumes that the imaging system or reconstruction algorithm is expected to be used repeatedly under similar conditions, and that there is an opportunity to simulate or experiment with the system to determine its long-term average performance. One then chooses the regularization parameter, α in the present case, that gives the best long-term performance. In our approach, performance is specified in terms of how well an observer of the images performs the task of interest. There is no controversy over such an approach to experimental design when one is in the repeated-use mode and has the opportunity to study average performance [10].

In the Bayesian perspective, one usually does not have the luxury of long-term experience. One has only the present data set and one's principles, i.e., that the only rational approach to such limited-data-set problems is by way of probability theory. In the Gull and Skilling approach, the regularization parameter is determined from the current data set by maximizing the posterior probability of α. In the "classic" implementation of MEMSYS3 this gives rise to the relationship

$$\chi^2 + G = N \ . \tag{2}$$

The interpretation of this expression is as follows. The aimed-for value of the misfit, χ^2, plus a parameter G, matches the number of independent measurements, N. The quantity G is referred to as the number of "good" measurements and is determined with α in a self-consistent way from the number of significant eigenvalues of a weighted version of the matrix $\mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}$. The weighting will be given in Section 10 below. Here, $\mathbf{H}$ is the system response matrix that characterizes the forward problem, $\mathbf{\Sigma}$ is the covariance matrix of the measurements, and the value of α is used in the determination of the significance of the eigenvalues.

Setting $G = 0$ in Eq. (2) yields the early "historic" implementation of MaxEnt where the target value of χ^2 is the number of independent measurements. Skilling and Gull [9] point out that the historic version is based on the frequentist consideration that the expected value of χ^2 over the ensemble of experiments is N. They

and other investigators realized that the historic approach often led to underfitting of the data. In the classic version, the higher the quality of the measurements in terms of the parameter G, the closer one is permitted to fit the data. The classic solution seems intrinsically reasonable, even setting aside its assumptions and subtlety.

5. Decision-theoretic measures of task performance

When studying binary tasks, for example decisions of "normal" vs "abnormal" or patent vs occluded artery, one is able to use the accepted, in fact now required, standard of image assessment in the medical imaging community, namely, the curve of true-positive fraction vs false-positive fraction of responses in the binary classification. This curve is referred to as the receiver operating characteristic (ROC) curve. We follow the usual approach of taking the area under the ROC curve, A_z, as a summary figure of merit for our studies. This measure is equivalent to the true-positive fraction averaged over all false-positive fractions. It may also be obtained as the percent correct in a two-alternative forced-choice experiment. It is convenient to use an inverse error function to convert from A_z to the detectability index, d_a, which may be thought of as the integrated signal strength in units of the standard deviation of its underlying noise distribution (a decision-theoretic signal-to-noise ratio). The uncertainty in the determination of d_a due to the use of a finite number of image samples is usually estimated from the number of image samples and the sampling statistics of the binomial distribution. See [5-7] for further details.

6. Results: Detection Task

The detection task we studied is an idealization of a low-contrast lesion-detection task in a background of measurement noise and artifacts generated by the presence of high-contrast structures. In Fig. 1a we show one realization of a phantom for generating images of this class. Ten randomly placed low-contrast "lesions" can be seen as faint disks. These are the lesions or disks to be detected. Ten randomly placed high-contrast disks or lesions, which serve as the major source of reconstruction artifacts, are easily seen. In Figs. 1b through 1e we show the results of reconstructing images over a range of values of α. At the highest value of α the reconstructions are smooth; at the lowest value the reconstruction is pointillist in texture. In the limit of very small α the reconstruction may be thought of as the maximum likelihood reconstruction with a positivity constraint.

In Fig. 2 we give results for d_a for three classes of observers. Among the classical and Bayesian observers, those derived from the likelihood function and those derived from the ratio of posterior probabilities performed similarly [6,7]; the best average results are shown and are labelled "machine observer". Results for machine, neural-net, and human observer show a similar falloff from the maximum level of performance at low values of α to the weakest level of performance at high values of α. The "classic" implementation of MEMSYS3 yields a value of α that corresponds to a point on the shoulder of the performance curves near the plateau.

Figure 1. (a) One realization of a scene used for generating images for the detection task. (b) through (e) - MEMSYS3 reconstructions with $\alpha = 0.002, 0.21, 1.8$, and 20.

Figure 2. Decision-theoretic signal-to-noise ratio (d_a) for performance of the detection task by machine, back-propagation neural network, and human observers. The total number of independent image trials for a given observer is the value N shown. The neural network was trained on $\frac{N}{2}$ and tested on the other $\frac{N}{2}$. The error bars represent $\pm 1\sigma$.

7. Results: Discrimination Task

We selected a discrimination task that is an elaboration of the Rayleigh task of discriminating between a single star or object and a doublet. This task also serves as an idealization of the task of determining whether a blood vessel is unobstructed or is narrowed. In Fig. 3a we show one realization of a phantom for generating images of this class. There are eight cigar-shaped singlets and eight doublets. Each class of objects serves as a source of reconstruction artifact generation for the detection of the other. In Figs. 3b through 3e we show the results of reconstructing images for various values of α.

In Fig. 4 we give results for d_a for the three classes of observers. These results all show a similar maximum of performance in the neighborhood of $\alpha = 1.0$, with a fall-off from this maximum for smaller and larger values of α. It is of interest that the "classic" value of α occurs close to the position of the peak of these curves. The small difference between the classical or Bayesian machine and the neural-net observers is within the error bars.

8. Summary from frequentist perspective

These results reinforce and extend the remarkable similarity in performance noted in earlier work among all three classes of observers for a given task. This is satisfying because an original goal of these investigations was to find visual-like machine observers, i.e., algorithms that performed similarly to human observers.

More important for our present purposes, however, is the comparison between performance on different tasks. Comparing Figs. 2 and 4 we see that the α-dependence of the task performance for the detection task is *qualitatively different* from the α-dependence of the task performance for the Rayleigh-like discrimination task for all three classes of observers. (Similar differences on a related problem have been observed by Abbey and Barrett [11].) From the frequentist perspective, then, we can say that the problem of selecting the optimal regularization parameter has a task-dependent character. This task-dependence is not addressed in the Bayesian formulations of the regularization problem. The knowledge of the task required of an imaging system or image reconstruction algorithm, however, is a form of prior information that is suitable for inclusion in a proper Bayesian formulation. We are not yet able to present such a Bayesian formulation. We are, however, able to analyze the structure of the present frequentist treatment with a view toward an ultimate Bayesian solution. The remainder of this paper summarizes our present understanding of the relevant issues.

9. The importance of eigenvectors

The *eigenvalues* of the matrix $\mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}$ were seen in Sec. 4 to play an important role in the Bayesian solution to the regularization problem. We contend, however, that the *eigenvectors* are the relevant quantities when task performance is brought into consideration. We briefly review our experience with a broad class of tasks.

Figure 3. (a) One realization of a scene used for generating images for the Rayleigh discrimination task. (b) through (e) - MEMSYS3 reconstructions with $\alpha = 0.05$, 0.6, 4.03, and 19.51.

Figure 4. Decision-theoretic signal-to-noise ratio (d_a) for performance of the discrimination task by machine, back-propagation neural network, and human observers. Otherwise, as in Fig. 2.

Hanson [12] considered a wide class of problems where an ideal observer would suppress the low spatial-frequency components of an image. Images in which the background is inhomogeneous (variable, lumpy, etc.) are members of this class because inclusion of the low-frequency components would only lead to the accumulation of irrelevant noise. It is easy to show that the task of determining the separation of two lesions, the task of determining whether a stenosis (a vessel narrowing) is present, and many other Rayleigh-like tasks are also members of this class. Their performance not only requires no low spatial-frequency information, it is even impeded by the presence of such information. A task representative of some linear shift-invariant members of this class was analyzed by Myers et al. [13]. They showed that the optimal linear filter for the task is a bandpass filter that maximizes the detection of Fourier-domain eigenvectors within a task-dependent band of frequencies; it suppresses the detection of eigenvectors outside this band so as to maximize noise rejection in irrelevant bands. Optimal signal detection here does not depend on any measure of eigenvalue number. It depends on optimal detection of the eigenvectors of interest. This feature of the problem is ignored in all Bayesian approaches that we are aware of.

In particular, the allowable stopping points of the MEMSYS3 algorithm in its search for an optimal α proceed from very large values where, in effect, only low spatial frequencies are reconstructed, to very small values where, in effect, low, intermediate, and high spatial frequencies are reconstructed. However, what is needed for Rayleigh-like tasks is a means to concentrate on a mid-band of frequencies. Allowing very low and very high spatial frequencies into the reconstruction will swamp all of the visual-like image observers (in ours and related work) with irrelevant noise.

The eigenvector issue is not unique to the imaging problem. The issue should be treated in neural-net design and in countless other signal and image processing problems where principles for optimal order determination or optimal stopping parameters are sought.

10. Toward a solution

We end with a sketch of a solution to this problem. The solution will require some fundamentals from statistical decision theory.

The frequentist figure of merit for ideal detection of a difference signal $\Delta\mathbf{f}$ (a *vector)* by an imaging system whose measurements are characterized by the matrix $\mathbf{\Sigma}_m^{-1} = \mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}$, in the case where there is no appreciable artifactual noise and the measurement noise is additive and Gaussian, is a detection theoretic signal-to-noise ratio (SNR) given by

$$d_a^2 = \Delta\mathbf{f}^t\mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}\Delta\mathbf{f} = \Delta\mathbf{f}^t\mathbf{\Sigma}_m^{-1}\Delta\mathbf{f} \; . \tag{3}$$

(See, e.g., [14, 15].) This figure of merit may be recognized as the Hotelling trace; it is proportional to the Mahalanobis distance between the two classes whose mean difference image is $\Delta\mathbf{f}$. It may be generalized for a Gaussian ensemble of signals with known covariance matrix $\mathbf{\Sigma}_p$ by replacing $\mathbf{\Sigma}_m$ with $[\mathbf{\Sigma}_m^{-1} + \mathbf{\Sigma}_p^{-1}]^{-1}$.

In the present paper the mean difference signal for the detection task is a template with a profile given by that of the expected lesion; in the discrimination task it is a template with a profile given by the expected difference of the mean signals of the two classes. The observer mask that will achieve the optimal figure of merit in the limit of many tomographic views is the prewhitening matched filter given by $\Delta\mathbf{f}^t\mathbf{H}^t\mathbf{\Sigma}^{-1}$. This filter selects out from the data only the vectors (eigenvectors or singular vectors) required for its task. It discards all others. The present paper is a demonstration of a frequentist search for optimal linear and nonlinear observer decision functions (as well as regularization parameters) for the sparse-view version of this problem.

Some of the machinery for evaluating and optimizing such figures of merit already exists in the powerful MEMSYS packages. There, Bayesian figures of merit play a central role. For example, the matrix designated $\mathbf{A}$ that generates the eigenvalues contributing to G has structure similar to Eq. (3):

$$\mathbf{A} = [\hat{\mathbf{f}}^{\frac{1}{2}}]^t\mathbf{H}^t\mathbf{\Sigma}^{-1}\mathbf{H}[\hat{\mathbf{f}}^{\frac{1}{2}}] \quad . \tag{4}$$

Here, $[\hat{\mathbf{f}}]$ is a diagonal *matrix* containing the current estimate of the object; $[\hat{\mathbf{f}}^{\frac{1}{2}}]$ is the corresponding square-root matrix. The number of significant eigenvalues of $\mathbf{A}$, on a scale determined by α, is the parameter G referred to earlier. In fact, $\alpha\mathbf{A}^{-1}$ is the MEMSYS analog of the quotient $\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_m$ that determines the degree of regularization in Gaussian MAP estimation [4].

A fundamental difference between Eqs. (3) and (4) can be noted in the first and last factors of these expressions. In the case of Eq. (3) these factors are frequentist or long-term average quantities that specify the task. In the case of the Bayesian result in Eq. (4) they are estimates from the present data set alone. From the point of view of the present paper, optimal reconstruction must address the problem of preferentially visualizing the eigenvectors required for the task, in the spirit of the optimal filter behind Eq. (3); it is not sufficient to count the eigenvalues of a regularization matrix such as $\mathbf{A}$ in Eq. (4). The details remain to be worked out.

Acknowledgements

Over many years the authors have enjoyed discussions and collaborations on the present and related issues with Prof. H.H. Barrett and colleagues at the University of Arizona.

References

1. K. Rossmann, "Image quality and patient exposure," *Current Problems in Radiology*, **2**(2), pp. 1-34, Year Book Medical Publishers, Chicago, 1972.
2. D.G. Brown, M.F. Insana, and M. Tapiovaara, "Detection performance of the ideal decision function and its McLaurin expansion: Signal position unknown," J. Acoust. Soc. Am. **97**, pp. 379-398, 1995.
3. R.F. Wagner, K.J. Myers, A.E. Burgess, D.G. Brown, and M.J. Tapiovaara, "Maximum a posteriori detection: Figures of merit for detection under uncertainty," Proc. SPIE **1232**, pp. 195-204, (Bellingham WA) 1990.
4. A.D. Whalen, *Detection of Signals in Noise,* (Electrical Science Series) Academic Press, New York, 1971.

5. K.M. Hanson, "Making binary decisions based on the posterior probability distribution associated with tomographic reconstructions," in *Maximum Entropy and Bayesian Methods,* C.R. Smith, G.J. Erickson, and P.O. Neudorfer, eds., pp 313-326, Kluwer Academic, Dordrecht, 1992.
6. K.J. Myers, R.F. Wagner, and K.M. Hanson, "Binary task performance on images reconstructed using MEMSYS3: Comparison of machine and human observers," in *Maximum Entropy and Bayesian Methods*, A. Mohammad-Djafari and G. Demoment, eds., pp. 415-421, Kluwer Academic, Dordrecht, 1993.
7. K.M. Hanson and K.J. Myers, "Rayleigh task performance as a method to evaluate image reconstruction algorithms," in *Maximum Entropy and Bayesian Methods,* W.T. Grandy Jr., and L.H. Schick, eds., pp. 303-312, Kluwer Academic, The Netherlands, 1991.
8. S.F. Gull and J. Skilling, *Quantified Maximum Entropy "MEMSYS 3" Users' Manual,* Maximum Entropy Data Consultants Ltd., Royston England, 1989.
9. J. Skilling and S.F. Gull, "Bayesian maximum entropy image reconstruction," in *Spatial Statistics and Imaging,* A. Possolo, ed., pp. 341-367, Institute of Mathematical Statistics Lecture Notes-Monograph Series **20**, IMS, Hayward CA, 1991.
10. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis,* Springer-Verlag, New York, 1985.
11. C.K. Abbey and H.H. Barrett, "Linear iterative reconstruction algorithms: Study of observer performance," in *Information Processing in Medical Imaging,* Y. Bizais et al. eds., pp. 65-76, Kluwer Academic, The Netherlands, 1995.
12. K.M. Hanson, "Variations in task and the ideal observer," Proc. SPIE **419**, pp. 60-67, 1983.
13. K.J. Myers, J.P. Rolland, H.H. Barrett, and R.F. Wagner, "Aperture optimization for emission imaging: Effect of a spatially varying background," J. Opt. Soc. Am. **A7**, pp. 1279-1293, 1990.
14. H.H. Barrett, "Objective assessment of image quality: Effects of quantum noise and object variability," J. Opt. Soc. Am. **A7**, pp. 1266-1278, 1990.
15. H.H. Barrett, J.L. Denny, R.F. Wagner, and K.J. Myers, "Objective assessment of image quality. II. Fisher information, Fourier crosstalk and figures of merit for task performance," J. Opt. Soc. Am. **A12**, pp. 834-852, 1995.